

DOCUMENT RESUME

ED 338 711

TM 017 587

AUTHOR de-la-Torre, Roberto; Vispoel, Walter P.
TITLE The Development and Evaluation of a Computerized Adaptive Testing System.
PUB DATE Apr 91
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Computer Simulation; Computer Software Development; Computer Software Evaluation; Educational Assessment; Elementary Secondary Education; Listening Comprehension Tests; Mastery Tests; Microcomputers; Online Systems; Questionnaires; *Scoring; Student Attitudes; *Test Construction

ABSTRACT

The development and preliminary evaluation of the Computerized Adaptive Testing System (CATSYS), a new testing package for IBM-compatible microcomputers, are described. CATSYS can be used to administer and score operational adaptive tests or to conduct on-line computer simulation studies. The package incorporates several innovative features, including skip and review options, content balancing, item eligibility flagging, mastery testing, self-adapted testing, and listening items. Two hundred and five examinees took tests produced by CATSYS (items were from the Iowa Tests of General Educational Development) and completed questionnaires about the test. Results indicate that the examinees: (1) preferred CATSYS-produced tests to paper-and-pencil alternatives; (2) found CATSYS-produced tests clear and simple to take; and (3) appreciated review, skip, and self-adaptive options. Results from small-scale computer simulations (responses to 400 adaptive tests, 4 for each examinee) indicate that the accuracy of ability estimates produced by CATSYS is comparable or superior to that of estimates provided by the BILOG and MicroCAT programs. Two tables present study data. There are 26 references. Three appendices contain examples of main menus used in CATSYS, a summary of questionnaire responses concerning examinees' reactions to tests produced by CATSYS, and examples of CATSYS score reports. (Author/SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED338711

The Development and Evaluation of A Computerized Adaptive Testing System

Roberto de-la-Torre and Walter P. Vispoel

The University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

WALTER VISPOEL

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) "

Paper Presented at the Annual Meeting of
The National Council on Measurement in Education
Chicago, Illinois
April, 1991

TM017587

Abstract

The purpose of this paper is to describe the development and preliminary evaluation of CATSYS (Computerized Adaptive Testing System)--a new adaptive and conventional testing package for IBM-compatible microcomputers. CATSYS can be used to administer and score operational adaptive tests or to conduct on-line computer simulation studies. This package incorporates several innovative features including: skip and review options, content balancing, item eligibility flagging, mastery testing, self-adapted testing, and listening items. Two-hundred-and-five examinees took tests produced by CATSYS and completed questionnaires about their opinions of it. Results indicated that the examinees a) preferred CATSYS-produced tests to paper and pencil alternatives, b) found CATSYS-produced tests clear and simple to take, and c) appreciated CATSYS's review, skip, and self-adaptive options. Results from small scale computer simulation studies indicated that the accuracy of ability estimates provided by CATSYS were comparable or superior to those provided by BILOGTM and MicroCATTM.

Background and Purpose

One of the most recent advances in measurement technology is the development of computerized adaptive tests. Research assessing the feasibility of adaptive tests indicates that adaptive tests require considerably fewer items to yield reliabilities and validities comparable or superior to those of conventional tests (McBride & Martin, 1983; Moreno, Wetzel, McBride & Weiss, 1984; Urry, 1977; Weiss, 1982). These promising results have led to the development of the first operational adaptive tests, which presently include adaptive versions of a) standardized aptitude tests, such as the Armed Services Aptitude Test Battery (Allfred & Green, 1984; Sympton, Weiss & Ree, 1982) and the Differential Aptitude Tests (McBride, 1986), b) college placement tests (Stocking, 1987), and c) achievement tests for elementary and high school students (cf. Rudner, 1989). In addition, the WICAT Systems Corporation and the Assessment Systems Corporation have created adaptive testing packages to enable users to construct their own adaptive tests.

The purpose of this paper is to describe a new adaptive testing package available for IBM-compatible microcomputers. The "Computerized Adaptive Testing System" (CATSYS) was developed as a research tool to support both theoretical and applied research on computerized adaptive testing. CATSYS can be used to conduct simulation studies on adaptive testing as well as to administer and score operational adaptive tests. In the following sections, the major components of this package are described, along with the results from preliminary studies evaluating a) the precision of ability estimates provided by the package and b) examinees' attitudes about the features of tests generated by the package.

Statistical Features of CATSYS

One of the important features of CATSYS is its flexibility in producing different kinds of adaptive and conventional tests. The various statistical options available in CATSYS are summarized under six major headings: 1) ability estimation methods, 2) item selection strategies, 3) test termination criteria, 4) item selection constraints, 5) general testing strategies, and 6) measurement models.

1) **ABILITY ESTIMATION PROCEDURES:** CATSYS is able to conduct the ability estimation stage of an adaptive testing strategy using four different statistical procedures: a) Maximum Likelihood Estimation (MLE); b) Owen's Bayesian adaptive testing strategy (Owen, 1969, 1970, 1975); c) Bayesian Modal (MAP); and d) EAP Algorithm (Bock and Mislevy, 1982). To avoid the problem of all right or all wrong answers, in which case a MLE solution does not exist, CATSYS uses Owen's Bayesian procedure until the examinee answers at least one item correctly and one item incorrectly.

2) **ITEM SELECTION STRATEGIES:** CATSYS has been designed to allow the test developer to select items from the available pool in several ways: a) Item difficulty; b) Match M to the current ability estimate (Hulin, Drasgow and Parsons, 1983); c) Maximum Information Selection; d) Minimum Posterior Variance (Owen, 1969, 1975); e) Random Item Selection; f) Fixed Branching (e.g. Stratadaptive Testing Strategy; Weiss, 1973); g) Examinee selected items (e.g. Self-Adapted tests; Rocklin and O'Donnell, 1987; Rocklin, 1989); and h) User-defined item sequences.

3) **TEST TERMINATION CRITERIA:** CATSYS' available options for terminating an adaptive test are: a) Precision level required by the test developer, as measured by the standard error of estimate; b) Fixed number of items to administer; c) Maximum number of items to administer; and d) Time constraints imposed by the test developer (as measured by time elapsed between the beginning and end of the test.)

4) **ITEM SELECTION CONSTRAINTS:** CATSYS incorporates some of the most recent suggestions for improving the validity of adaptive tests by imposing additional constraints on item selection. (Kingsbury and Zara, 1989). One of these optional constraints is "Content Balancing". CATSYS uses a procedure outlined by Kingsbury and Zara (1989), that assures that examinees respond to a specific number or proportion of items from each of the user-defined content strata. Another constraint is the "Eligibility of Items". CATSYS keeps track of encounters between examinees and items to avoid, if requested, presenting examinees with flagged items, such as those presented on prior tests, those with conflicting content, and so forth.

5) **GENERAL TESTING STRATEGIES** Tests may be designed for either mastery/nonmastery decisions or ability estimation (survey testing). For mastery decision-oriented adaptive testing, CATSYS uses the procedures outlined by Weiss and Kingsbury (1984).

6) **MEASUREMENT MODELS:** CATSYS will perform Classical Test Theory analyses based on number correct or formula scores, and Item Response Theory analyses based on the one- two- and three-parameter logistic models.

Computer Program Description

CATSYS consists of five main subprograms:

- 1) **CREATEST:** This is a menu-based program that defines the parameters (ability estimation procedure, item selection strategy, stopping rule, etc.) desired in implementing an adaptive or conventional test. This subprogram also is used to define the kind of reports and data analyses to perform after the simulated or actual adaptive tests have been administered. (See Appendix A for examples of some of these menus.)
- 2) **SIMULATE:** This program is designed to simulate adaptive tests according to specifications provided in the CREATEST stage.
- 3) **ADAPTIVE:** This program builds and administers actual adaptive tests, according to the specifications developed in the CREATEST stage.
- 4) **REPORT:** This program creates the reports defined in the CREATEST subprogram the summarize the performance of simulated or actual examinees on the test of interest...
- 5) **ANALYSIS:** This program performs several kinds of item analysis, according to the specifications provided in the CREATEST subprogram.

Evaluation Data For CATSYS.

The plan for evaluating CATSYS performance is divided into four major categories: 1) Test development and administration, including test assembly, item presentation, screen control, and user friendliness; 2) Speed and precision of the test scoring procedures; 3) Breadth and clarity of reporting system; and 4) Portability to different computer systems. Information collected to date for each of these categories follows.

Test Development and Administration

The test development procedures in CATSYS are very straightforward. The test developer simply responds to a series of menus (see Appendix A for examples) that identify the type of test to be developed, and the locations of computer files and interfaces. Three components are needed to implement a given test: 1) an item parameter file, 2) an item text/content file, and 3) the appropriate interface to administer the test. (An interface is not needed for computer simulation studies.).

We have successfully implemented operational computerized tests using CATSYS in two recent studies. In the first study, items from the Iowa Tests of Educational Development (Iowa Testing Programs, 1987) were used to create five computerized vocabulary tests: 1) a fixed-item test without review, 2) a fixed-item test with review, 3) an adaptive test without review, 4) an adaptive test with review, and 5) a self-adapted test. These tests were administered to 205 examinees with approximately 41 examinees responded to each type of test. After completing a given computerized test, each examinee responded to a questionnaire assessing his or her reactions to the test. Selected items from this questionnaire, along with corresponding item means and percentages of examinees responding to each item alternative are given in Appendix B. As can be seen from the appendix, examinees reactions were very favorable concerning the visual displays from CATSYS (items 1,2,3,16,17,18) and the skip, review, and self-adapted options it can provide (items 4,5,7,21,22,23,26,27,28,48). Moreover, as evident from items 34 to 45, examinees consistently preferred the CATSYS tests to paper and pencil tests measuring the same skills.

In the second study in which CATSYS was used, three computerized tests of tonal memory (the ability to remember melodies) were implemented: 1) a fixed-item test without review, 2) an adaptive test without review, and 3) a self-adapted test. (Note that no review options were used in this study since such options are rarely included on any kind of listening test). The development of these tests was a major breakthrough because they require aurally- rather than visually-administered items. CATSYS is the only current microcomputerized adaptive testing system that permits test developers to construct such items. This was accomplished rather easily by including appropriate PASCAL commands in the item text file that produce music directly through the computer speaker. These tests were administered to 60 college marching band students along with questionnaires similar to those used in the first study. The data about examinees reactions to the adaptive music tests, however, are not available at this time.

Precision of the Ability Estimates

Two small scale simulation studies were conducted comparing the precision of ability estimates derived from CATSYS and two other packages, MicroCATTM (Assessment Systems Corporation, 1989), and BILOGTM (Mislevy and Bock, 1982).

Purpose of Study 1: The purpose of the first simulation study was to evaluate the precision of the various ability estimation procedures available in CATSYS. (The other packages were not included in this study.) The first step in implementing this study was to create a pool of quality items. This was accomplished by selecting 200 items from five forms of the vocabulary subtests from the Iowa Tests of Educational Development (Iowa Testing Programs, 1987). These 200 items were selected from an original pool of 275 items in the following way

1) The 275 unique vocabulary items from the five original paper and pencil forms of the ITED were equated and calibrated using LOGIST V (Wingersky, Barton, and Lord, 1982). In total, these calibrations were based on responses to 10,000 examinees--five randomly equivalent samples of 2000 examinees responded to each test form.

2) Items were ordered from easiest to most difficult, and categorized into 8 difficulty strata. Within each stratum, items were ordered by discrimination from most to least discriminating.

3) The final pool was constructed by successively selecting the most discriminating items from each stratum. This process was repeated until 200 items were chosen and resulted in the construction of a pool containing highly discriminating items at all ability levels.

Generation of Item Response Data: Item responses for the 200 item pool were generated in the following way:

1) Let U represent a random variable, with $U=1$ if the item is answered correctly and $U=0$ otherwise. Then, the probability of a correct response to item i is given by:

$$P(U=1|\text{Theta}_k, a_i, b_i, c_i) = P_i(\text{Theta}_k) \quad (i=1, \dots, n)$$

Where Theta_k represents the true ability level for examinee k , and a_i , b_i , and c_i represent the discrimination, difficulty, and guessing parameters for item i ($i=1, \dots, n$). Once the mathematical form of $P_i(\text{Theta}_k)$ is given, and the values of Theta_k , a_i , b_i , and c_i are known, then the value of $P_i(\text{Theta}_k)$ can be determined. The model chosen to generate $P_i(\text{Theta}_k)$ was the Three-Parameter Logistic Model.

2) To generate known ability levels, 1000 random values from a standard normal distribution were generated. Each value so generated was used as the true ability for each examinee.

3) For a particular examinee, responses to the 200 items were generated by calculating $P_i(\text{Theta}_k)$ for $i=1, \dots, n$. A random number, R_i , from a (0,1) uniform distribution was generated. If $P_i(\text{Theta}_k) > R_i$, then the response was considered to be correct. Otherwise, the response was considered to be incorrect. The resulting values were considered as the true responses that each examinee would provide to the 200 items in the pool if they were administered in normal conditions.

Generation of Simulated Adaptive Tests: To evaluate CATSYS' accuracy in recovering the examinees' true abilities, the administration of four types of adaptive tests were simulated. To convenience sake, the response vectors for only the first 100 of the 1000 simulated examinee response patterns, in point 2 above, were used in this analysis. Each adaptive test consisted of 40 items, used a maximum information item selection procedure, and was administered using each of the four ability estimation strategies [Maximum Likelihood Estimation (MLE), Owen's Bayesian, Bayesian Modal (MAP), and the EAP Algorithm]. In total then, responses to 400 adaptive tests were simulated--four adaptive tests for each examinee.

Each adaptive test was administered as if it was an actual adaptive test, according to the particular test specifications and testing strategy. That is, items were selected from an actual item pool and each item was administered and scored using the item responses

previously simulated for each examinee. Ability estimates were obtained and recorded after each item was administered.

Indices of Precision: Four indices were used to evaluate the precision of ability estimates provided by CATSYS.

- 1) Bias. Mean difference between true and estimated ability across examinees.
- 2) Inaccuracy. Mean absolute difference between true and estimated ability across examinees.
- 3) RMSE. Square root of mean squared difference between true and estimated ability across examinees.
- 4) Fidelity. Correlation between true and estimated ability.

Results. Table 1 provides a summary of the results for the precision of ability estimation for each of the four procedures. As shown in Table 1, the indices of precision for the ability estimates provided by CATSYS were highly accurate (e.g., fidelity correlations of .969 and higher), and--with a couple of minor exceptions--were very consistent across ability estimation procedures. Specifically, the Bayesian Modal procedure (BMP) provided slightly larger bias than the other procedures and the Owen's procedure provided slightly less accurate ability estimates (as evidenced by slightly higher Inaccuracy and RMSE index values).

Table 1
Ability Estimation Precision Results for CATSYS

	BIAS	ACCURACY	RMSE	FIDELITY
EAP	-0.0191	0.1756	0.2387	0.970
OWEN	-0.0161	0.1812	0.2428	0.969
MLE	-0.0108	0.1759	0.2369	0.970
MAP	-0.0348	0.1728	0.2358	0.972

Purpose of Study 2: The purpose of the second study was to compare the precision of the ability estimates provided by CATSYS to the precision of ability estimates provided by BILOGTM and MicroCATTM.

Generation of Item Response Data and Simulated Tests: Item response data for Study 2 were generated using procedures similar to those used in Study 1. As was the case in Study 1, 100 ability levels were generated from a standard normal distribution. Based on these values, 100 vectors of item responses were generated, but for a 50-fixed-item test rather than a 40-item adaptive test. A fixed-item test was used in Study 2 because there is no straightforward way to simulate an adaptive test using MicroCATTM. The fixed-item test used for this simulation was built by selecting the first 50 items from the item pool generated for Study 2. This resulted in a fixed-item test containing highly discriminating items over a wide range of ability levels. The average "a" value was 1.55 (standard deviation = .43, minimum = .89, maximum = 2.79); the average "b" value was .03 (standard deviation = 1.17, minimum = -2.06, maximum = 2.89); and the average "c" value was .18 (standard deviation = .04, minimum = .10, maximum = .28). The simulated response vectors for the 100 examinees were scored using the appropriate sub-programs for the three computer programs being compared. Each program produced EAP, MLE, and MAP estimates.

Indices of Precision: The indices of precision were the same as those used to compare CATSYS ability estimates to each other in Study 2--Bias, Inaccuracy, RMSE, and Fidelity--plus one additional index, the Mean Variance Error of Estimate. Consistent with Mislevy and Bock (1982), indices for the Mean Variance Error of Estimate for each estimation procedure were defined as follows: EAP, the variance of the final posterior distribution; MLE, the negative reciprocal of the second derivative log likelihood at the final ability estimate; and MAP, the negative reciprocal of the curvature at the mode of the final posterior distribution (These indices were available only for BILOGTM and CATSYS; MicroCATTM does not provide standard errors or variance errors for theta scores)

Results: A summary of the results for the precision of ability estimates obtained by the three computer programs is given in Table 2. As can be seen from the table, CATSYS provided similar or, in some cases, superior levels of measurement precision compared to the other programs. Results for CATSYS and MicroCATTM were essentially identical. However, the MLE and MAP procedures in CATSYS and MicroCATTM produced somewhat better precision than the MLE and MAP procedures in BILOGTM in terms of Bias, Inaccuracy, and RMSE. Variance Error indices also were lower for CATSYS than for BILOGTM.

Table 2.
Ability Estimation Precision Results for CATSYS,
BILOGTM, and MicroCATTM

	Estimation Procedure		
	EAP	MLE	MAP
Bias			
CATSYS	0.0185	-0.0021	-0.0034
BILOG	0.0189	-0.0290	-0.0286
MicroCat	0.0185	-0.0019	-0.0032
Inaccuracy			
CATSYS	0.1928	0.1962	0.1938
BILOG	0.1923	0.1991	0.2068
MicroCat	0.1928	0.1963	0.1939
RMSE			
CATSYS	0.2483	0.2514	0.2483
BILOG	0.2428	0.2562	0.2667
MicroCat	0.2484	0.2514	0.2483
Fidelity			
CATSYS	0.9649	0.9642	0.9654
BILOG	0.9647	0.9628	0.9641
MicroCat	0.9649	0.9642	0.9653
Mean Variance Error of Estimate			
CATSYS	0.0517	0.0562	0.0463
BILOG	0.0515	0.1708	0.1294

Breadth of Scoring System

At this time, the scores provided by CATSYS and other adaptive testing programs have not been compared systematically. The current default displays provided by CATSYS for survey and mastery tests are given in Appendix C. For survey tests, the following scores are provided: number correct, percent correct, percentile rank, stanine, t-score, estimated item pool number correct, estimated item pool percent correct, theta, standard error of theta, 95% confidence interval for theta, and reliability of theta. For mastery tests, the same scores are provided along with a mastery/ nonmastery score. Item by item information for each examinee is saved in a separate file. This file contains the examinee's name and/or identification number, and the following information for each item: 1) the test item number, 2) pool item number, 3) examinee's selected response, 4) the correct response, 5) the examinee's current theta score, and 6) the current standard error of the given theta score.

Portability of the System

CATSYS has run successfully on a variety of IBM and IBM-compatible microcomputers with 8088, 286, and 386 microprocessors including IBMTM XTs and ATs, LEADING EDGETM, EPSENTM and COMPAQTM. A MACINTOSHTM version of the program also is being developed.

Summary and Conclusions

Although CATSYS clearly is still in development, the evaluation data obtained to date is encouraging. Examinees seem to prefer CATSYS-produced tests to paper and pencil alternatives, find CATSYS-produced tests clear and simple to take, and appreciate CATSYS' review, skip, and self-adaptive options. Preliminary data comparing the precision of ability estimates indicates that CATSYS provides results comparable or superior to those provided by BILOGTM and MicroCATTM. The capabilities of CATSYS to administer aurally-produced items and to perform computer simulation studies set it apart from other programs. These advantages are expected to make CATSYS a particularly useful tool for many researchers and test developers.

References

- Allfred, L. A., & Green, B. F. (1984). Analysis of experimental CAT ASVAB test data. Baltimore: John Hopkins University, Department of Psychology.
- Assessment Systems Corporation (1989). User's manual for the Microcat. St. Paul, MN: Author.
- Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 37, 431-444.
- Hulin, J. L., Drasgow, F., & Parsons, C. K. (1983). Item Response Theory: Application to Psychological Measurement. Homewood, IL.: Dow Jones-Irwing.
- Iowa Testing Programs (1987). Manual for teachers, administrators, and counselors: Iowa Tests of Educational Development. Iowa City, IA: The University of Iowa.
- Kingsbury, G. G. & Zara, A. R. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. Applied Measurement in Education, 2(4), 359-375.
- McBride, J. R. (1986). A computerized adaptive edition of the differential aptitude test. (ERIC Document Reproduction Service No. ED 285918). Psychological Corporation.
- McBride, J. R. (1977). Some Properties of a Bayesian Adaptive Ability Testing Strategy. Applied Psychological Measurement, 1(1).
- McBride, J. R., & Martin, J. T. (1983). Reliability and Validity of adaptive ability tests in a military setting. In D. J. Weiss (Eds.). New Horizons in testing: Latent trait theory and computerized adaptive testing. New York: Academic Press.
- Mislevy, R. & Bock, R. D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logist models. Mooresville, IN: Scientific Software.
- Moreno, K. R., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. Applied Psychological Measurement, 8(2), 155-163.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal responses in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.
- Owen, R. J. (1970). Bayesian sequential design and analysis of dichotomous experiments with special reference to mental testing. Unpublished paper. Department of Statistics. University of Michigan.
- Owen, R. J. (1969). A Bayesian Approach to Tailored Testing. Research Bulletin, 69-92. Princeton, N.J.: Educational Testing Service.
- Rocklin, T. (1989, March). Individual differences in item selection in computerized self-adapted testing. Paper presented at the meeting of the American Educational Research Association, San Francisco.

- Rocklin, T. & O'Donnell, A. M. (1987). Self-Adapted Testing: A performance-improving variant of computerized adaptive testing. Journal of Educational Psychology, 79, 315-319.
- Rudner, L.M. (1989). Notes from ERIC/TM. Educational Measurement: Issues and Practice, 8(4), 25-26.
- Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 263-277.
- Sympson, J. B., Weiss, D. J., & Ree, M. (1982). Predictive validity of conventional and adaptive tests in an Air Force training environment (AF HRL-TR-81-40). Brooks Air Force Base, TX. USAF Human Resources Laboratory.
- Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14(2), 181-196.
- Vale, C. D., & Weiss, D. J. (1975). A simulation study of stradaptive ability testing. Research Report 75-6. Department of Psychology. Minneapolis, MN.: Universidad of Minnesota.
- Vale, C. D. y Weiss, D. J. (1974) A Study of Computer-Administered Stradaptive Ability Testing. Research Report 75-4. Department of Psychology. Minneapolis, MN.: University of Minnesota.
- Weiss, D. J., & Kingsbury (1984). Application of Computerized Adaptive Testing to Educational Problems. Journal of Educational Measurement, 21(4), 361-375.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.
- Weiss, D. J. (1974). Strategies of Adaptive Ability Measurement. Research Report 74-6. Department of Psychology. Minneapolis, MN.: Universidad of Minnesota.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. LOGIST V, version 1.0. Princeton, NJ: Educational Testing Service.

APPENDIX A

Examples of Main Menus Used in CATSYS

DIRECTIONS

THIS PROGRAM ALLOWS YOU TO DEFINE THE PARAMETERS NEEDED TO RUN THE MAIN PROGRAM.

FIRST, THE PROGRAM ASKS YOU FOR THE NAME OF THE FILE WHICH CONTAINS THE ITEM PARAMETERS. AFTER PERFORMING THE SYSTEM INITIALIZATION, THE PROGRAM WILL LEAD YOU THROUGH SEVERAL TASKS TO DEFINE A COMPLETE TESTING STRATEGY.

THE PROGRAM ACCEPTS INPUT IN THREE DIFFERENT WAYS:

1) MENUS

Use UP and DOWN arrows to place the inlighted cursor on the desired option, then press ENTER to select the option.

2) YES/NO QUESTIONS

Just type "Y" or "N" to the prompt.

3) TYPED INPUT

The prompt will indicate the kind OF input requested. It can be a Real value, an Integer value or a String.

Type the requested input using the keyboard and then press the ENTER key.

|| PRESS ANY KEY TO CONTINUE ||

TESTING STRATEGY DEFINITION

|| CREATE TEST SPECIFICATIONS ||

||

|| ADMINISTRATION/SELECTION CONSTRAINTS ||

||

|| CREATE FILES ||

||

|| QUIT ||

-Cursor

ITEM SELECTION/ADMINISTRATION CONSTRAINTS

REVIEW TEST ALLOWED (Y/N)?N

FEEDBACK AFTER EACH ITEM (Y/N)?N

SKIPPING ITEMS ALLOWED (Y/N)?N

IMMEDIATE RESPONSE REQUIRED (Y/N)?Y

CONTENT BALANCING PROVIDED (Y/N)?

FILE INFORMATION EDITING

FILE CONTENT	FILE NAME AND LOCATION
ITEM PARAMATER (OUTPUT)	C:\VOCFILES\REALDATA\VOCAPARA.REC
FINAL ABILITY ESTIMATES	C:\VOCFILES\REALDATA\VOCAPARA.RES
CONTROL FILE	C:\VOCFILES\REALDATA\PRUEBA1.CON
INFORMATION TABLE	C:\VOCFILES\REALDATA\VOCAPARA.INF
LIKELIHOODS	NOT REQUIRED

↓-Cursor Esc-Exit

OPTIONS FOR THE RUN

ADAPTIVE TEST	FULL ADAPTIVE TESTING STRATEGY
TYPE OF ESTIMATION	SURVEY
METHODS OF ABILITY ESTIMATION	BAYESIAN OWEN STRATEGY
METHODS OF ITEM SELECTION	MAXIMUM INFORMATION
FILES STATUS	DEFINED
STOPPING CRITERION	FIXED TEST LENGTH
MIN. NUMBER OF ITEMS TO ADMINISTER	10
MAX. NUMBER OF ITEMS TO ADMINISTER	10
NUMBER OF ITEMS TO ADMINISTER	10

ARE THESE OPTIONS CORRECT (Y/N)?

RESULTS OF ADAPTIVE TESTING SESSION

||A. DISPLAY INFORMATION FOR QUESTIONS ||
||B. DISPLAY INFORMATION ON TEST SCORES ||
||
||Q. QUIT TESTING SESSION ||
||

-Cursor

APPENDIX B

Summary of Examinees' Reactions to Tests Produced by CATSYS

For items 1 to 7, the numbers given represent the percentage of examinees circling each alternative. Item means appear under the text for each item

		1 %	2 %	3 %	4 %	5 %	6 %	
1. Computer screen text readability Mean = 1.24	Sharp	81	15	2	1	0	0	Fuzzy
2. Test directions Mean = 1.21	Clear	85	10	3	1	0	0	Unclear
3. Practice items Mean = 1.66	Helpful	68	11	14	1	4	1	Not Helpful
4. Permitting examinees to go over the items more than once Mean = 2.25	Important to include as part of the test	46	20	12	8	8	5	Not important to include as part of the test
5. Permitting examinees to skip items and return to them later Mean = 2.98	Important to include as part of the test	29	17	16	14	12	12	Not important to include as part of the test
6. Permitting examinees to review the directions after the test has started Mean = 2.48	Important to include as part of the test	42	18	18	5	9	9	Not important to include as part of the test
7. Practice items Mean = 3.09	Important to include as part of the test	23	10	26	11	12	12	Not important to include as part of the test

For items 14 to 20, the numbers given represent the percentage of examinees circling each alternative. Item means appear under the text for each item

		About Right								
		1 %	2 %	3 %	4 %	5 %	6 %	7 %		
14. Test length Mean = 4.24	Too Short	0	0	4	74	15	4	2	Too Long	
15. Test difficulty Mean = 3.96	Too Easy	1	6	11	63	15	4	0	Too Hard	
16. Computer screen brightness Mean = 4.13	Too Dim	0	0	0	89	8	1	1	Too Bright	
17. Computer screen text size Mean = 3.92	Too Small	1	0	5	92	1	0	0	Too Large	
18. Computer screen text spacing Mean = 3.95	Too crowded	0	0	0	95	4	0	0	Too spread out	
19. Difficulty of practice items Mean = 2.95	Too Easy	19	20	16	40	5	0	0	Too Difficult	
20. Number of practice items Mean = 3.76	Too few	5	6	9	73	5	1	1	Too Many	

For items 21 to 30, the numbers given for each alternative represent the percentage of examinees marking that alternative. Item means appear under the text for each item.

	Greatly improve my score %	Moderately improve my score %	Slightly improve my score %	Not affect my score %	Slightly decrease my score %	Moderately decrease my score %	Greatly reduce my score %
21. Going over the items more than once Mean = 3.26	5	12	42	36	3	1	0
22. Skipping items and returning to them later Mean = 3.36	5	11	32	48	2	1	1
23. Permitting examinees to select the difficulty level of items Mean = 2.92	20	19	22	33	5	1	1
24. Reviewing the directions after the test had started Mean = 3.51	4	9	22	66	0	0	0
25. Including practice items Mean = 3.10	10	13	37	39	0	0	0

	Greatly reduce my anxiety level %	Moderately reduce my anxiety level %	Slightly reduce my anxiety level %	Not affect my anxiety level %	Slightly increase my anxiety level %	Moderately increase my anxiety level %	Greatly increase my anxiety level %
26. Going over the items more than once Mean = 3.37	7	13	32	37	8	2	0
27. Skipping items and returning to them later Mean = 3.40	8	13	25	39	12	1	0
28. Permitting examinees to select the difficulty level of items Mean = 2.92	22	16	20	33	5	2	1
29. Reviewing the directions after the test had started Mean = 3.43	5	9	27	54	3	1	0
30. Including practice items Mean = 2.85	15	18	35	28	3	0	0

For items 31 to 43 the numbers given for each alternative represent the percentage of examinees marking that alternative. Chi-square tests were for differences in the proportion of examinees favoring paper and pencil versus computerized tests produced by CATSYS. ($df = 1$ for each test).

	Paper and Pencil Test	Computerized Test	No difference between the two %	Chi-Square/ Significance
	%	%	%	
31. Which type of text display is easier to read?	10	52	38	57.80/.000
32. Which type of test creates the most anxiety?	38	22	41	8.53/.003
33. Which type of test do you prefer taking?	25	57	17	25.30/.000
34. On which type of test would you perform better?	20	32	48	5.43/.020
35. To which type of test is it easier to respond?	18	60	22	46.81/.000
36. For which type of test is it easier to record answers?	19	60	21	21.93/.000
37. Which type of test is more comfortable to take?	22	49	30	20.54/.000
38. Which type of test is more enjoyable to take?	7	75	17	112.39/.000
39. Which type of test yields fairer results?	15	20	66	1.47/.225
40. On which type of test would you be most likely to receive the same score if you took it a second time?	21	32	46	4.48/.034
41. Which type of test more accurately measures your knowledge of the subject?	15	15	70	0.00/1.00
42. Which type of test is <u>less</u> fatiguing?	13	62	24	64.06/.000
43. Which type of test have you taken more often?	99	1	1	190.18/.000

The percentage of examinees responding to options a and b for item 48 are given below. The Chi-square value for the difference between these proportions is 30.19, $df = 1$, $p < .00001$

48. Which type of test would you prefer taking?	%
a. one that permits you to choose the difficulty level of items.	69
b. one that does NOT permit you to choose the difficulty level of items.	31

APPENDIX C

Examples of Final Score Report Provided By CATSYS

SUMMARY RESULTS

NUMBER OF QUESTIONS ANSWERED	10	ITEM POOL SCORE	185.736
NUMBER OF QUESTIONS RIGHT	6	ITEM POOL PERCENTAGE RIGHT	67.54%
NUMBER OF QUESTIONS WRONG	4	THETA ESTIMATE	0.611
PERCENTAGE OF RIGHT ANSWERS	60.00%	STANDARD ERROR OF ESTIMATE:	0.277
PERCENTILE RANK: 73	INTERVAL: (53, 88)	INTERVAL FOR TRUE THETA:	
T-SCORE	: 56	INTERVAL: (51, 62)	$0.068 \leq \theta \leq 1.154$
STANINE	: 6	INTERVAL: (5, 7)	ESTIMATED RELIABILITY: 0.925

|| PRESS ANY KEY TO CONTINUE ||